An Efficient Algorithm to Automated Discovery of Interesting Positive and Negative Association Rules

Ahmed Abdul-Wahab Al-Opahi
Department of Computer Science
Faculty of Computer Sciences and Information Systems,
Thamar University
Thamar, Yemen

Basheer Mohamad Al-Maqaleh
Department of Information Technology
Faculty of Computer Sciences and Information Systems,
Thamar University
Thamar, Yemen

***Abstract*—Association Rule mining is very efficient technique for finding strong relation between correlated data. The correlation of data gives meaning full extraction process. For the discovering frequent items and the mining of positive rules, a variety of algorithms are used such as Apriori algorithm and tree based algorithm. But these algorithms do not consider negation occurrence of the attribute in them and also these rules are not in infrequent form. The discovery of infrequent itemsets is far more difficult than their counterparts, that is, frequent itemsets. These problems include infrequent itemsets discovery and generation of interest negative association rules, and their huge number as compared with positive association rules. The interesting discovery of association rules is an important and active area within data mining research. In this paper, an efficient algorithm is proposed for discovering interesting positive and negative association rules from frequent and infrequent items. The experimental results show the usefulness and effectiveness of the proposed algorithm.**

***Keywords*—Association rule mining; negative rule and positive rules; frequent and infrequent pattern set; apriori algorithm**

I. INTRODUCTION

Association rules (ARs), a branch of data mining, have been studied successfully and extensively in many application domains including market basket analysis, intrusion detection, diagnosis decisions support, and telecommunications. However, the discovery of associations in an efficient way has been a major focus of the data mining research community [1–2].

Traditionally, the association rule mining algorithms target the extraction of frequent features (itemsets) ie, features boasting high frequency in a transactional database. However, many important itemsets with low support (i.e. infrequent) are ignored by these algorithms.

These infrequent itemsets, despite their low support, can produce potentially important negative association rules (NARs) with high confidences, which are not observable among frequent data items. Therefore, discovery of potential negative association rules is important to build a reliable decision support system. The research in this paper extends discovery of positive as well as negative association rules of the forms A→¬B (or ¬A→B, ¬A→¬B), and so on.

The researchers target three major problems in association rule mining:

a) effectively extracting positive and negative association rules from real-life datasets.

b) extracting negative association rules from the frequent and infrequent itemsets.

c) the extraction of positive association rules from infrequent itemsets.

The rest of this paper is organized as follows. In the second section, related work on association rule mining. In third section, description of interesting positive and negative association rules is presented. The fourth section, the proposed algorithm for discovering interesting positive and negative association rules is described. Experimental results are shown in fifth section. Conclusion and future work are presented in the sixth section.

II. RELATED WORK

A standard association rule is a rule of the form A→ B, where A and B are frequent itemsets in a transaction database and A∩B=Ø. This rule can be interpreted as "if itemset A is true of an instance in a database, so is itesmset B true of the same instance", with a certain level of significance as measured by two indicators, support and confidence. Rule support and confidence are two measures of rule interesting. What if we have a rule such as A→¬B, which says that the presence of A in a transaction implies that B is highly unlikely to be present in the same transaction. Rules of the form A→¬B are called negative rules. Negative rules indicate that the presence of some itemsets will imply the absence of other itemsets in the same transactions [3].

Support-confidence framework for discovering association rules. The validity of an association rule has been based on two measures: the support; the percentage of transactions of the database containing both A and B; and the confidence; the percentage of the transactions in which B occurs relatively only to those transactions in which also A occurs [4].

Investigated the efficient mechanism of identifying positive and negative associations among frequent and infrequent itemsets using state-of the-art data mining technology is presented in [5].

Genetic algorithm GA for mining interesting rules from dataset has proved to generate more accurate results when compared to other formal methods available. The fitness function used in GA evaluates the quality of each rule [6].

Efficacy contemplations for discovering interesting rules from frequent itemsets are suggested in [7-8].

A framework for fuzzy rules that extends the interesting measures for their validation from the crisp to the fuzzy case is presented in [9].

A fuzzy approach for mining association rules using the crisp methodology that involves the absent items is proposed in [10].

Another study introduced to extract interesting association rules from infrequent items by weighting the database "the weight of database must be determined and used the frequent items to discover the infrequent items" [11].

An interesting association rules mining algorithm is proposed to integrate Rule Interestingness measure during the process of mining frequent itemsets, which generates interesting frequent itemsets [12].

Traditional association rules algorithms mostly concentrate on positive association rules. Also, they generate a large number of rules, many of which are redundant and not interesting to the users. The interestingness measures can be used an effective way to filter and then reduce the number of discovered association rules. Based on that, a unified framework is proposed for mining a complete set of interesting positive and negative association rules from both frequent and infrequent itemsets simultaneously

III. Description of Positive and Negative Association Rules

Discovering association rules between items in large databases is a frequent task in knowledge discovery in database KDD. The purpose of this task is to discover hidden relations between items of sale transactions. This later is also known as the market basket database. An example of such a relation might be that 90% of customers that purchase bread and diaper also purchase milk.

TABLE I. Database with 5 Transactions

Transaction	Items
1	Bread,Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Let D be a database of transactions. Each transaction consists of a transaction identifier and a set of items {i1,i2 , ...,in} selected from the universe I of all possible descriptive items. Let D be a database of transactions as shown in Table 1.

The items represents the customer database of sale transactions as a basket data. Each record in this database consists of items bought in a transaction. The problem is how it can be found some interesting (i.e. hidden) relations existing between the items in these transactions or some interesting rules that a manager (a user, a decider or a decision-maker) who owns this database can take some valuable decisions. Some rules derived from this database can {Coke}→{Milk},{Diaper}→{Beer},{Coke,Milk}→{Diaper}.

A positive association rule is an expression of the form: A→B . Each association rule is characterized by means of its support and its confidence defined as follows: Supp (A→B) =Number of transactions containing (AUB) / Total number of transactions. conf (A→B) =supp (A→B) / supp (A). From the above example, rule {Coke}→{Milk} has support 40% and confidence 100%. According to the above measures, the support measure can be considered as the percentage of database transactions for which (AUB) evaluates to be true. The confidence measure is understood to be the conditional probability of the consequent given the antecedent. Association rule mining essentially boils down to discovering all association rules having support and confidence above user-specified thresholds, minsup and minconf, for respectively the support and the confidence of the rules. For example, from the 100% confidence of the rule {Coke},{Diaper}→ {Milk}. It can be concluded that customers that purchase coke and diaper also purchase milk.

In the dataset, it exists other association rule: A→¬B, ¬A→B, ¬A→¬B. The rule A→¬B means the data objects which have itemsets A do not have the itemsets B. The rule ¬A → B means the data objects which do not have itemsets A have the itemsets B. The rule ¬A→¬B means the data objects which do not have itemsets A do not have the itemsets B. These rules can be called negative association rules. For the above example, from the 75% confidence of the rule {Bread}→ {¬Coke}. It can be concluded that customers that purchase bread will not also purchase coke. The rule A→B can be called positive association rule. In the existing paper researchers expressed their views on negative association rule in Basket Market database. It is negative association rule which is very useful to the market basket administrator to adjust the business decision making from the customers database. It resolves the lack of past which is only researching positive association rules. This makes the decision makers and access pattern is mined more objective and comprehensive. In order to calculate, the support and confidence for negative association, it can be computed the measures through those of positive rules.

1) *Supp (¬A) = 1-supp (A);*
2) *Supp (A U ¬B) = supp (A)-supp (A U B);*
3) *Supp (¬A U B) = supp (B) - supp (A U B);*
4) *Supp (¬A U ¬B) = 1- supp (A) - supp (B) + supp (A UB)*
5) *Conf (A → ¬B) = supp (A) - supp (A U B)/ 1-supp (A) =1- conf (A→B);*
6) *conf(¬A→B)= supp(A) - supp(A U B)/1-supp (A) = supp (B)- supp(A) * conf (A => B) /1 - supp(A)*
7) *conf(¬A→¬B)=1 - supp(A) - supp(B) + supp(A U B) = conf(¬A→B)/1 - supp (A)*
8) *Lift (A→B) =supp (A U B)/ supp (A)*supp (B)*

IV. THE PROPOSED ALGORITHM

The proposed algorithm for automated discovery of interesting positive and negative associations rules consist of two steps:

A. Finding all frequent and infrequent item sets in the database D.

B. Mining interesting association rules (both positive and negative) from the itemsets which we get in the first step.

The interesting measure (lift) has to be greater than one, expressing a positive dependency among the itemsets. The value of lift less than one will express a negative relationship among the itemsets. Figure 1. shows the proposed algorithm.

V. EXPERIEMENTAL RESULTS

The performance of the proposed algorithm on different datasets is demonstrated below and all the codes are implemented under C# language.

A. EXPERIMENT 1

Weather dataset is downloading from the UCI datasets repository. This dataset contains twelf items, fourteen transactions, and seventy words. It helps the researchers in weather forecasts. This datasets applied with varying minsupport and minconfidence values in table 2.We can see that the number of frequent itemsets decreases as we increase the minsupport value. However, a sharp increase in the number of infrequent itemsets can be observed. This can also be visualized in figure 2.

1 Generating all the Candidate k-itemsets(Ck) which its support>0
  1.1 if Ck>= minsupport
    Then frequent itemsets. Add(Ck)
  1.2 else if Ck<minsupport
    Then Infrequent itemsets. Add(Ck)
2 Generating all interesting association rules from frequent itemsets
  2.1 for each items I in frequent itemsets
  2.2 Generating the rules of the form A==>B
  2.3 If confidence (A==>B)>= minconfidence&& lift(A==>B)>=1
  Then output the rule(A==>B) as frequent positive association rules
3 Generating all the interesting association rules from infrequent itemsets
  3.1 for each items I in Infrequent itemsets
  3.2 Generating the rules of the form A==>B
  else
  Then output the rule(A==>B) as Infrequent positive association rules
  3.3 If confidence (A==>B)>= minconfidence&& lift(A==>B)>=1
  3.4 Generating the rules of the form(A==>~B) &&(~A==>B)&&(~A==>~B)
  3.5 if confidence(~A==>B)>= mincofidence&& Lift(~A==>B)>=1
  Then output the rule of form(~A==>B) as Infrequent negative rules
  3.6 if confidence(A==>~B)>=mincofidence&& Lift(A==>~B)>=1
  Then output the rule of form(A==>~B) as Infrequent negative rules
  3.7 if confidence(~A==>~B)>= mincofidence&& Lift(~A==>~B)>=1
  Then output the rule of form(~A==>~B) as Infrequent negative rules

Fig. 1. The proposed algorithm

TABLE II. TOTAL GENERATED FREQUENT AND INFREQUENT ITEMSETS USING DIFFERENT SUPPORT VALUES

Support	Frequent	Infrequent
0.1	104	136
0.15	42	198
0.20	42	198
0.25	22	218
0.3	11	229

Table 3. gives an account of the experimental results for different values of minimum support and minimum confidence. The liftF value has to be greater than one for a positive relationship between the itemsets; the resulting rule, however, may itself be positive or negative. The total number of positive rules and negative rules generated from both frequent and infrequent itemsets is given. Generates negative association rules of the A→¬B, ¬A→B, ¬A→¬B, which have greater confidence than the user defined threshold and lift greater than one, are extracted as negative association rules figure 3.

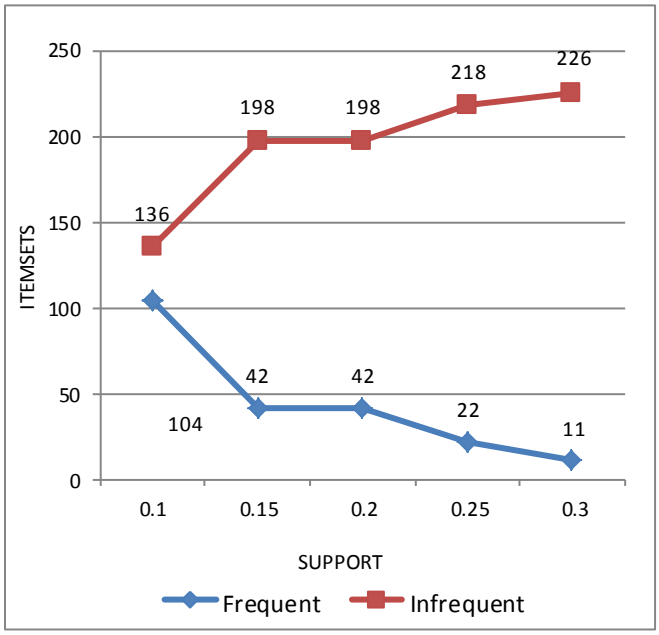

Fig. 2. Frequent and infrequent itemsets generated with varying minimum support values

TABLE III. INTERSITING POSITIVE AND NEGATIVE ASSOCIATION RULES USING VARYING SUPPORT AND CONFIDENCE VALUES WITH LIFT>1

Supp.	Conf.	PARs From Freq.	PARs From Infrq.	NARs From Freq.	NARs From Infrq.
0.1	0.5	166	351	55	137
0.1	0.7	54	193	15	39
0.15	0.5	35	468	8	134
0.15	0.7	11	236	3	51
0.25	0.5	12	491	0	105

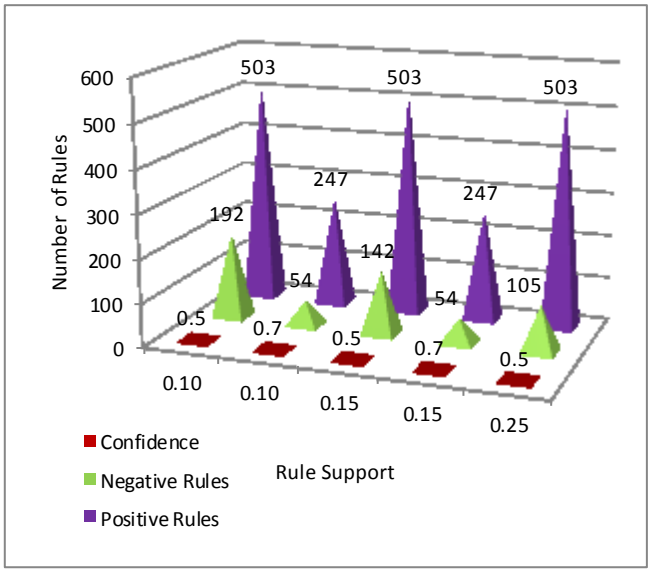

Fig. 3. Intersiting positive and negative association rules generated with varying minimum supports and confidence values

B. EXPERIEMENT 2

The Groceries dataset contains one month (30 days) of real-world point-of-sale transaction data from a typical local grocery outlet. The data set contains 9835 transactions , the items are aggregated to 169 categories and the total number of words 43367.The frequent and infrequent itemset generation using Apriori algorithm takes only an extra time as compared to the traditional frequent itemset finding using Apriorialgorithm.This is because each item's support is calculated for checking against the threshold support value to be classified as frequent and infrequent; therefore, we get the infrequent items in the same pass as we get frequent items. The proposed algorithm implemented for Groceries dataset to mine positive and negative from frequent and infrequent items with different parameters (minsupport,minconfidence,3 items length). Table 4. shows that the number of frequent itemsets decreases as it increase the minsupport value. However, a sharp increase in the number of infrequent itemsets can be observed. This can also be visualized in figure 4.

The total number of positive rules and negative rules generated from both frequent and infrequent itemsets which is given in Table 5. Generates negative association rules of the form A→¬B, ¬A→B, ¬A→¬B, which have greater confidence than the user defined threshold and lift greater than one, are extracted as negative association rules figure 5.

TABLE IV. TOTAL GENERATED FREQUENT AND INFREQUENT ITEMSETS USING DIFFERENT SUPPORT VALUES

Support	Frequent	Infrequent
0.00015	70858	78369
0.00025	44260	104966
0.0005	24172	125025
0.001	9969	139248
0.002	3812	145395

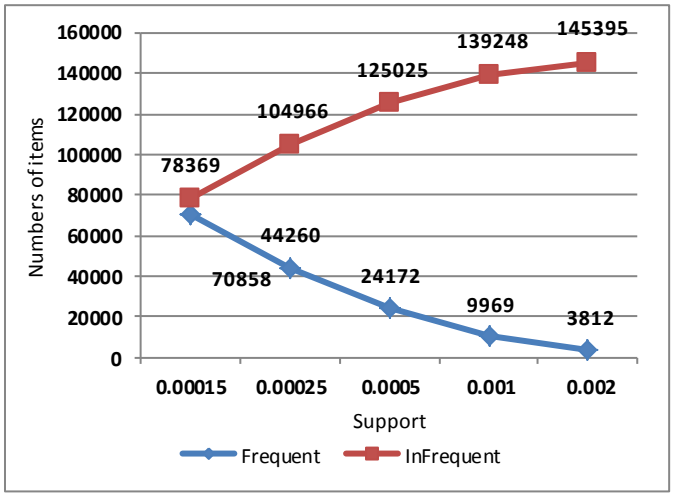

Fig. 4. Frequent and infrequent itemsets generated with varying minimum support

TABLE V. INTERSITING POSITIVE AND NEGATIVE ASSOCIATION RULES USING VARYING SUPPORT AND CONFIDENCE VALUES WITH LIFT>1

Supp.	Conf.	PARs From Freq.	PARs From Infrq.	NARs From Freq.	NARs From Infrq.
0.0005	50	3772	57496	2451	19190
0.001	50	1472	59835	881	19178
0.001	60	493	31514	881	19178
0.002	50	582	60725	330	15787
0.002	60	138	31869	330	15787

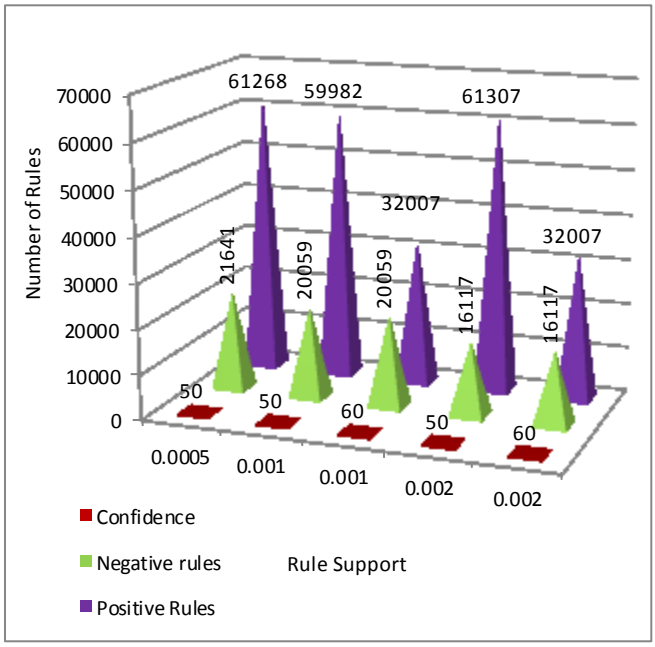

Fig. 5. Intersiting positive and negative association rules generated with varying minimum supports and confidence values

VI. Conclusion and Future Work

In this paper, a new algorithm to generate interesting positive and negative association rules from frequent and infrequent itemsets is proposed. Whereas, traditional association rules mining algorithms have focused on frequent items to generate positive association rules. The proposed algorithm integrates lift as interestingness measure during the process of mining rules. The experimental results have demonstrated that the proposed algorithm is efficient and promising. In future work the researchers will present improved algorithm by using different interestingness measures for mining association rules.

References

[1] F. H. AL-Zawaidah, Y. H. Jbara, and A. L. Marwan, "An Improved Algorithm for Mining Association Rules in Large Databases," Vol. 1, No. 7, 311-316, 2011

[2] H. H. O. Nasereddin, "Stream data mining," International Journal of Web Applications, vol. 1, no. 4, pp. 183–190, 2009.

[3] X. Wu, C. Zhang, and S. Zhang, "Efficient mining of bothpositive and negative association rules," ACM Transactions onInformation Systems, vol. 22, no. 3, pp. 381–405, 2004.

[4] R. Agrawal, T. Imieli´nski, and A. Swami, "Mining association rules between sets of items in large databases," ACM SIGMOD Record, vol. 22, no. 1, pp. 207–216, 1993.

[5] S. Mahmood, M. Shahbaz, and Z. Rehman, "Extraction ofpositive and negative association rules from text: a temporal approach," Pakistan Journal of Science, vol. 65,pp. 407–413, 2013.

[6] G. pKaur ,and S. Aggarwal, "A Survey of Genetic Algorithm for Association Rule Mining," International Journal of Computer Applications, vol.67, no.20, pp. 19-22, 2014.

[7] T. Slimani, A. Lazzez,"Efficient Analysis of Pattern and Association Rule Mining Approaches," International Journal of Information Technology and Computer Science, vol.6, no.3, pp.70-81, 2014.

[8] E. Duneja,A. K. Sachan,"A Survey on Frequent Itemset Mining with Association Rules,"International Journal of Computer Applications, vol.64, no.23, pp.7105-97202014.

[9] M. Delgado, M. D. Ruiz, D. S´anchez, and J. M. Serrano, "A formal model for mining fuzzy rules using the RL representation theory," Information Sciences, vol. 181, no. 23, pp. 5194–5213, 2011.

[10] M. Delgado, M. D. Ruiz, D. Sanchez, and J. M. Serrano, "A fuzzy rule mining approach involving absent items," in Proceedings of the 7th Conference of the European Society for Fuzzy Logic and Technology, pp. 275–282, Atlantis Press, 2011.

[11] L. Cagliero and P. Garza, "Infrequent Weighted Itemset Mining Using Frequent Pattern Growth," IEEE Transactions on know ledge and data engineering, vol.26, no.4, 2014.

[12] B. M. Al-Maqaleh and S. S. Ghalib,"Pushing Rule Interestingness Measure in Association Rules Mining," TUJNAS,Accepted, 2015.